ACH3657

Métodos Quantitativos para Avaliação de Políticas Públicas

Aula Teórica 02

Análise de Regressão

Alexandre Ribeiro Leichsenring

alexandre.leichsenring@usp.br

Introdução

- Modelo de regressão simples pode ser usado para estudar a relação entre duas variáveis
- Tem algumas limitações
- Apropriado como ferramenta empírica
- Aprender a interpretar o modelo de regressão simples: boa prática para estudar regressão múltipla

Definição do Modelo de Regressão Simples

- Premissa: y e x são duas variáveis e estamos interessados em explicar y em termos de x
- Ou: Estudar como y varia com variações em x

Exemplos

- y é a produção de soja; x é a quantidade de fertilizante
- y é o salário; x, anos de educação
- y é a taxa de criminalidade; x é o número de policiais

Questões

i Nunca há uma relação exata entre duas variáveis, como consideramos outros fatores que afetam y?

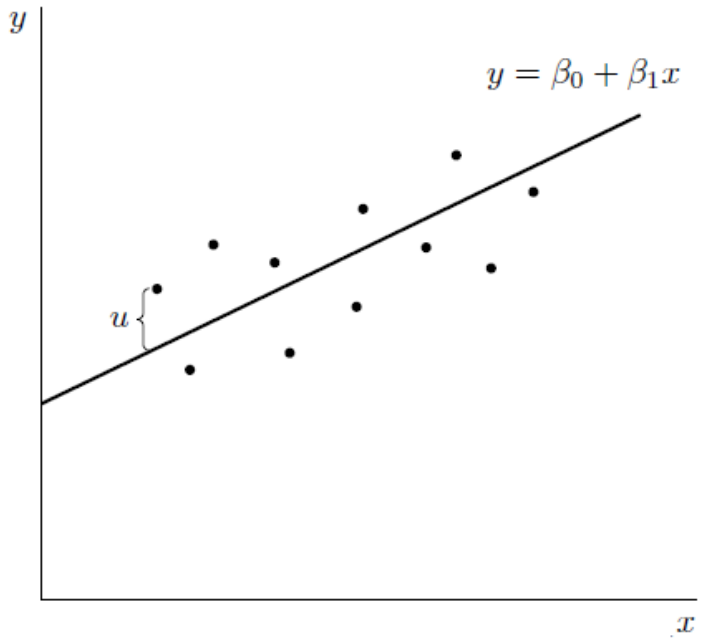
ii Qual a relação funcional entre x e y?

iii Como podemos estar certos de caputrar a relação *ceteris paribus* entre x e y?

- Uma equação simples é:

$$y = \beta_0 + \beta_1 x + u. \tag{1}$$

- A equação acima define o **modelo de regressão linear simples**



Terminologia para a Regressão Simples

y	x
Variável Dependente	Variável Independente
Variável Explicada	Variável Explicativa
Variável de Resposta	Variável de Controle
Variável Prevista	Variável Previsora
Regressando	Regressor

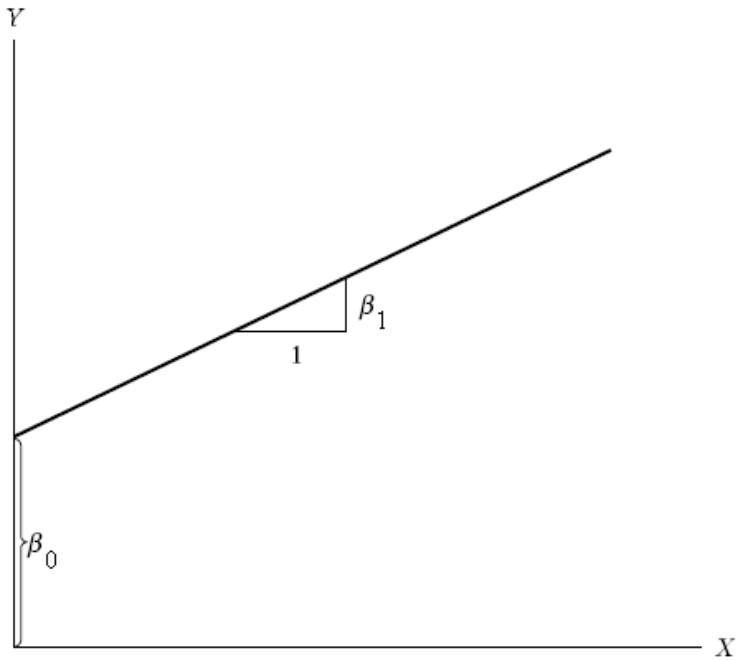
$$y = \beta_0 + \beta_1 x + u$$

- A variável u, chamada **termo de erro** ou **perturbação**, representa outros fatores, além, de x, que afetam y
- A regressão simples trata todos os fatores que afetam y, além de x, como **não-observados**
- Podemos pensar em u como o **não-observado**
- A equação também trata da relação funcional entre y e x
  - Se $\Delta u = 0$, então x tem um efeito linear sobre y;

$$\Delta u = 0 \Rightarrow \Delta y = \beta_1 \Delta x$$

- A variação em y é igual a β_1 multiplicado pela variação em x

- β_1 é o **parâmetro de inclinação** da relação entre y e x
- β_0 é o **parâmetro de intercepto** (raramente central em uma análise)



Exemplo: Produção de Soja e Fertilizantes

- Suponha que a produção de soja seja dada pelo modelo:

$$producao = \beta_0 + \beta_1 fertilizante + u$$

de modo que:

  - $y = producao$
  - $x = fertilizantes$

- O pesquisador agrícola está interessado no efeito dos fertilizantes sobre a produção, mantendo outros fatores fixos. (Efeito dado por β_1)
- O termo erro u contém fatores como qualidade da terra, chuva etc.
- O coeficiente β_1 mede o efeito dos fertilizantes sobre a produção, mantendo outros fatores fixos:

$$\Delta producao = \beta_1 \Delta fertilizante$$

Exemplo: Equação Simples do Salário

- Um modelo que relaciona o salário de uma pessoa à educação observada e outros fatores não-observados é

$$salario = \beta_0 + \beta_1 educ + u$$

- Se $salario$ é medido em reais por hora e $educ$ corresponde a anos de educação formal, β_1 mede a variação no salário-hora dado um ano a mais de educação, mantendo todos os outros fatores fixos
- Alguns desses fatores incluem experiência da força de trabalho, aptidão inata, permanência com o empregador atual, ética no trabalho, etc

- A linearidade do modelo implica que uma variação de uma unidade em x tem o mesmo efeito sobre y, independentemente do valor inicial de x
- Irreal em algumas aplicações
- Na equação salário x educação, podemos querer considerar retornos *crescentes*: o próximo ano de educação teria, em relação ao anterior, um efeito *maior* sobre os salários
- A questão mais difícil é saber se o modelo realmente nos permite tirar conclusões *ceteris paribus* sobre como x afeta y
- Acabamos de ver que β_1 mede o efeito de x sobre y, mantendo todos os outros fatores (em u) fixos
- Isso não encerra, entretanto, a questão da causalidade!
- Como podemos aprender algo sobre o efeito *ceteris paribus* de x sobre y, mantendo todos os outros fatores constantes, se estamos ignorando todos os outros fatores?!

- Veremos que para obter estimadores confiáveis de β_0 e β_1, precisaremos restringir a maneira como se relacionam o termo não-observável u e a variável explicativa x
- Sem tal restrição, não é possível estimar o efeito *ceteris paribus*, β_1.
- Antes, vamos supor o seguinte sobre u:

$$\mathbf{E}(u) = 0 \tag{2}$$

- A hipótese (2) não diz nada sobre a relação entre u e x
- Ela apenas faz uma afirmação sobre a distribuição dos fatores não-observáveis na população.

Covariância

Dados n pares de valores $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, observações de duas variáveis X e Y, chamamos de *covariância* entre X e Y a

$$\mathrm{cov}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

Coeficiente de Correlação

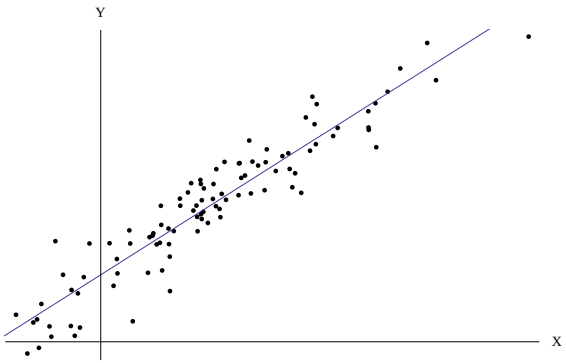
Chamamos de *coeficiente de correlação* entre X e Y a

$$\mathrm{corr}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\mathrm{dp}(X) \cdot \mathrm{dp}(Y)},$$

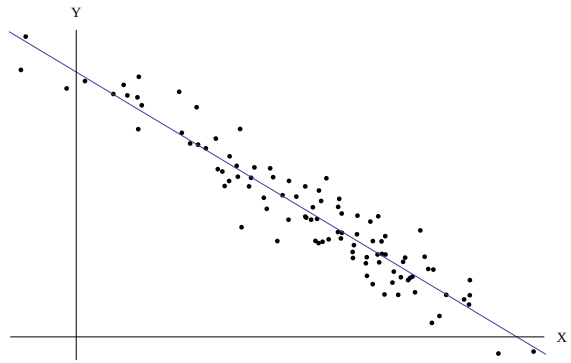
onde $\mathrm{dp}(X)$ = desvio padrão de X, e $\mathrm{dp}(Y)$ = desvio padrão de Y.

$$-1 \leq \mathrm{corr}(X, Y) \leq 1$$

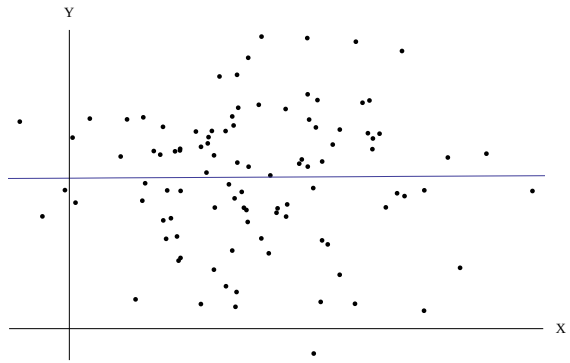
r = 0,94 (associação positiva forte)



r = −0,94 (associação negativa forte)



r = 0,07 (associação fraca)



- Vamos voltar à hipótese sobre a relação entre u e x
- Para qualquer valor de x, podemos obter o valor esperado (ou médio) de u para aquela fatia da população descrita pelo valor de x
- A hipótese crucial é que o valor médio de u não depende do valor de x. Podemos escrever isso como:

$$\mathbf{E}(u|x) = \mathbf{E}(u) = 0 \tag{3}$$

- A hipótese (3) diz que, para qualquer valor de x, a média dos fatores não-observáveis é a mesma e, portanto, deve igualar-se ao valor médio de u na população.
- A hipótese (3) é chamada **hipótese de média condicional zero**

- Vamos ver, por exemplo, o que a hipótese de média condicional zero acarreta no exemplo dos salários
- Para simplificar, vamos supor que u seja o mesmo que aptidão inata
- Então, (3) requer que o nível médio de aptidão seja o mesmo, independentemente dos anos de educação formal
- Por exemplo, se:
  - $\mathbf{E}(aptidao|educ = 8)$ representa a aptidão média para o grupo de todas as pessoas com oito anos de educação formal, e
  - $\mathbf{E}(aptidao|educ = 16)$ representa a aptidão média entre pessoas na população com 16 anos de educação formal
- Então (3) implica que essas médias devem ser as mesmas
- Se, por exemplo, entendemos que a aptidão média aumenta com os anos de educação formal, então (3) é falsa
  - Isso aconteceria se, em média, pessoas com maior aptidão escolhessem tornar-se mais escolarizadas

- Como não podemos observar aptidão inata, não temos um modo de saber se a aptidão média é ou não a mesma para todos os níveis de educação
- Essa é uma questão que devemos resolver antes de aplicar a análise de regressão simples
- No exemplo dos fertilizantes, se as quantidades de fertilizantes são escolhidas independentemente de outras características dos lotes, então (3) se sustentará
  - a qualidade média da terra não dependerá da quantidade de fertilizantes
- Entretanto, se mais fertilizantes forem usados em lotes de terra de melhor qualidade, então o valor esperado de u varia com o nível de fertilizantes, e (3) não se sustenta

Questão

Suponha que a nota de um exame final ($nota$) dependa da frequência às aulas ($freq$) e de fatores não-observados que afetam o desempenho dos estudantes (tal como a aptidão). Então:

$$nota = \beta_0 + \beta_1 freq + u$$

Em que situação você espera que esse modelo satisfaça a hipótese de média condicional zero?

- Considerando o valor esperado de

$$y = \beta_0 + \beta_1 x + u$$

condicionado a x are usando $\mathbf{E}(u|x) = 0$, obtém-se

$$\mathbf{E}(y|x) = \beta_0 + \beta_1 x$$

- A equação acima mostra que a função de regressão populacional, $\mathbf{E}(y|x)$, é uma função linear de x:
  - o aumento de uma unidade em x faz com que o valor esperado de y varie β_1

- Para qualquer valor dado de x, a distribuição de y está centrada em $\mathbf{E}(y|x)$:

$\mathbf{E}(y|x)$ como função linear de x